Artificial Intelligence: The Next Paradigm Shift in Medical Education

Cornelius A. James, MD
Erkin Ötleş, MS

9/14/2023

Objectives

- Define artificial intelligence (AI) and machine learning (ML)
- Describe the impact that AI/ML will have on health care
- Summarize the current state of AI/ML in medical education
- Provide a vision for AI/ML in medical education
- **Provoke thought and dialogue**

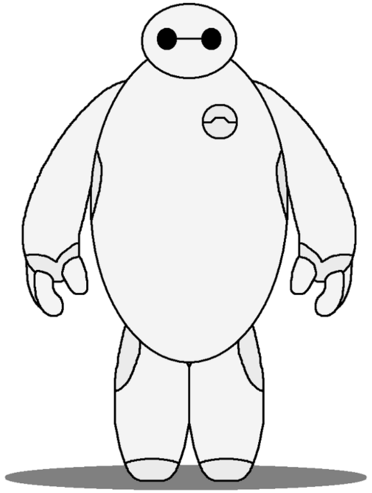
Disclosures

- Dr. James: none applicable

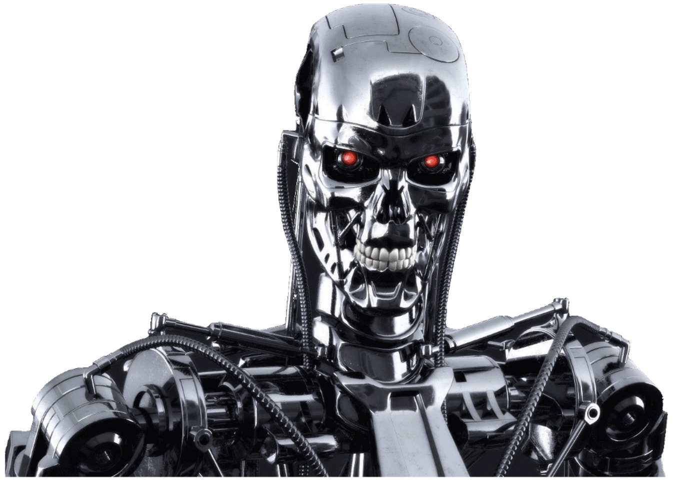
- Erkin: none directly related to today’s talk
  - Patent pending: AI prediction of health outcomes in patients with occupational injuries.
  - Small amount of IRA stock in various technology & healthcare companies.
  - Provide AI advising for several companies.

What comes to mind when you think about AI?

Acceptance | Annoyance | Avoidance



Hello! I am Baymax, your personal healthcare companion.



What is AI?

What is AI?

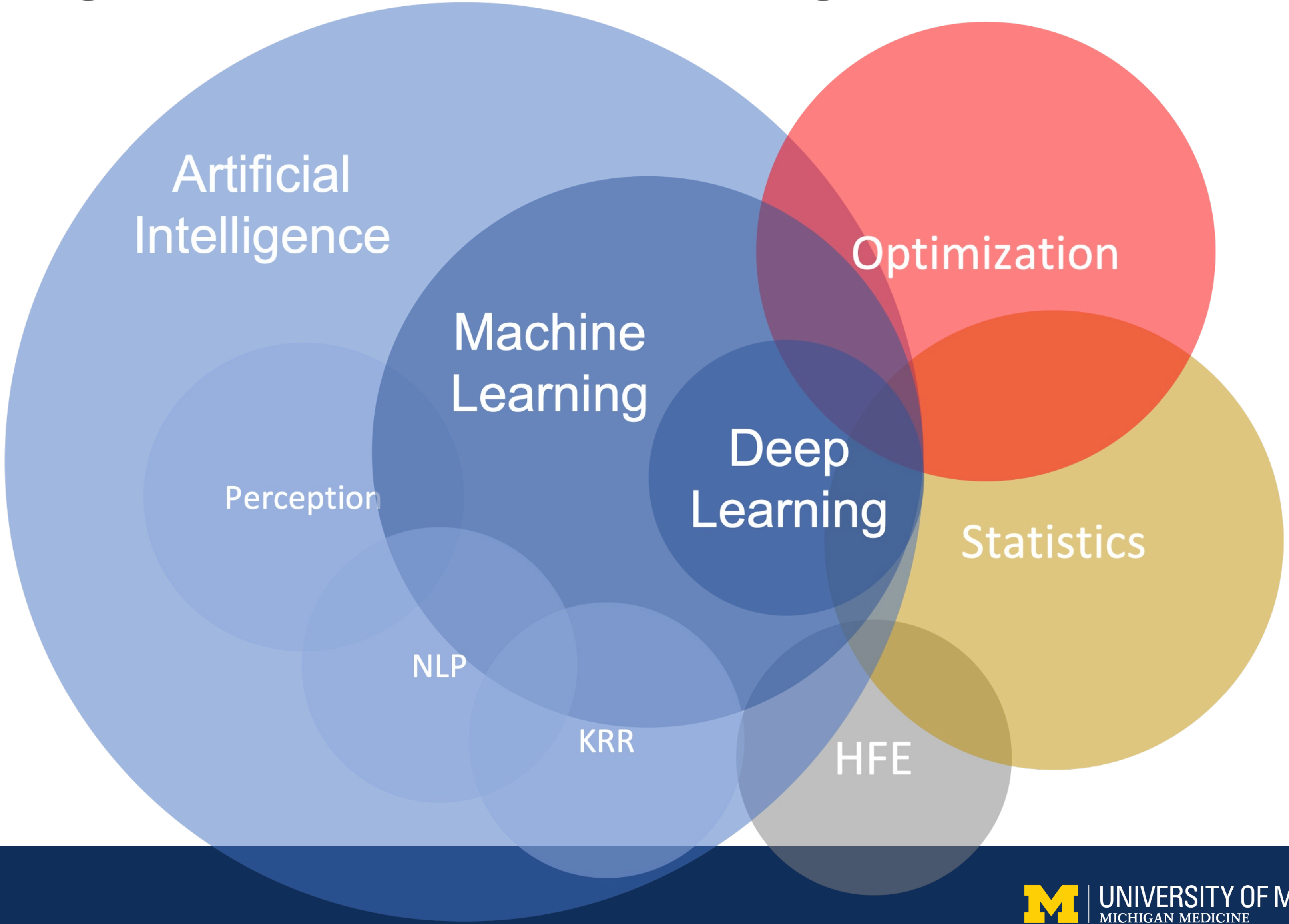
It is <u>not</u> magic.

First, some definitions

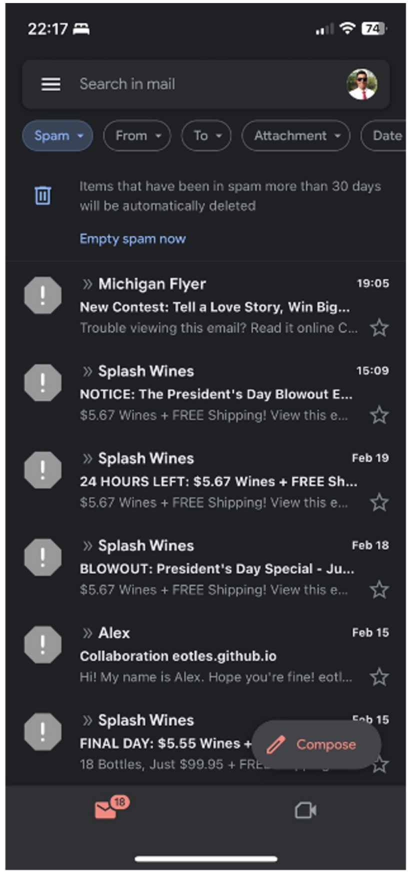
- **Artificial Intelligence (AI)**: *intelligence* (perceiving, synthesizing, and inferring information) demonstrated by machines

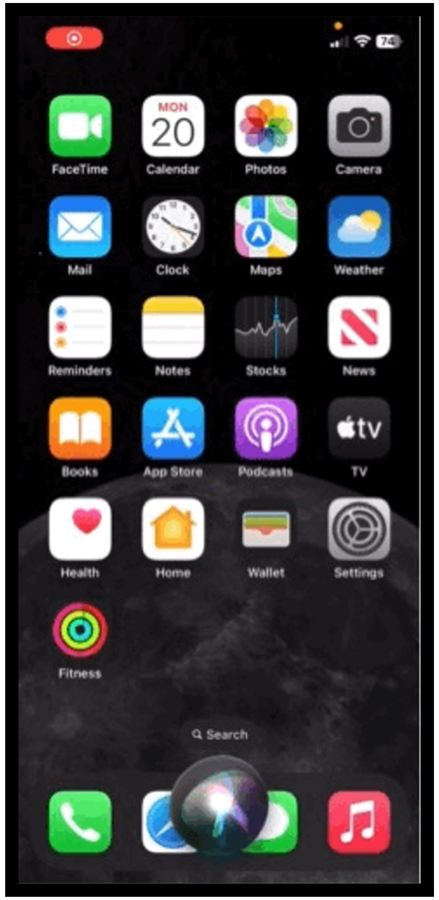
- **Machine Learning (ML)**: field of inquiry devoted to understanding and building methods that *learn* (use data to improve performance on a task).

Nesting and overlapping concepts



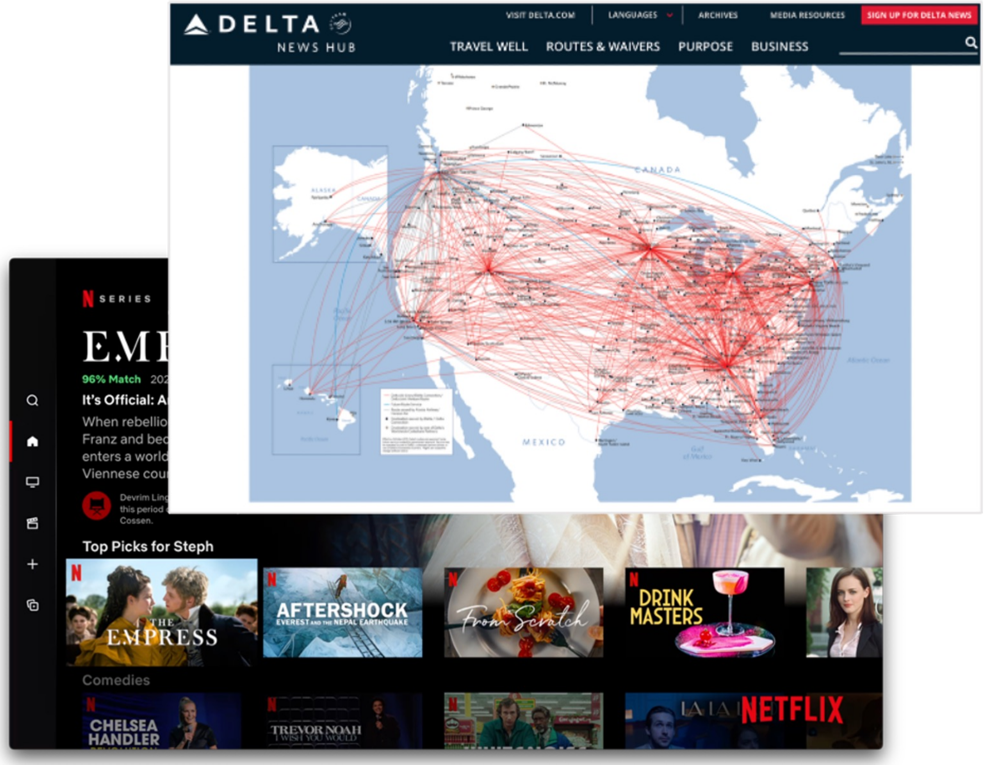
AI is ubiquitous in everyday life





Many industries depend on AI

- What routes should we fly?
- When should we service our planes?
- How should we price a product?
- What content should we serve?
- What products should we stock?



How does ChatGPT work?

ChatGPT = Chatbot + GPT3

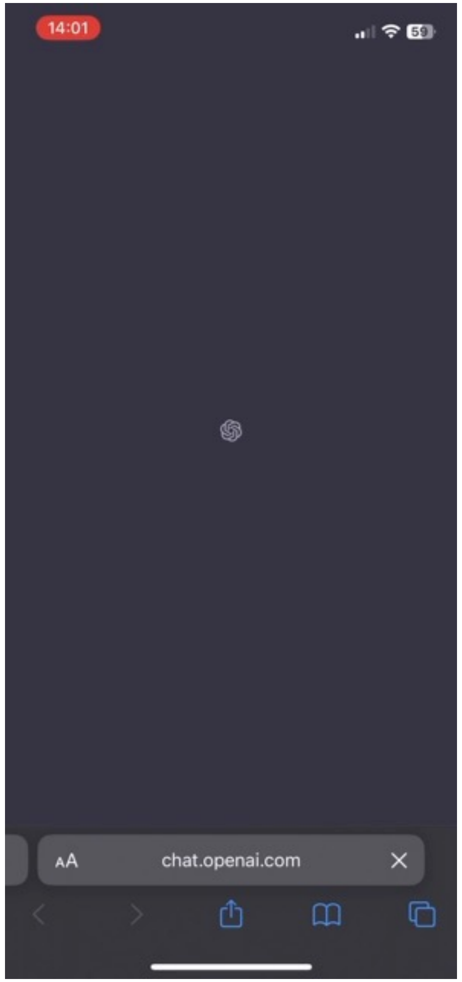
- Chatbot: developed by OpenAI
  mix of supervised & reinforcement learning

- GPT3: Generative Pre-trained Transformer 3
  type of **large language model** (fancy predictive text)

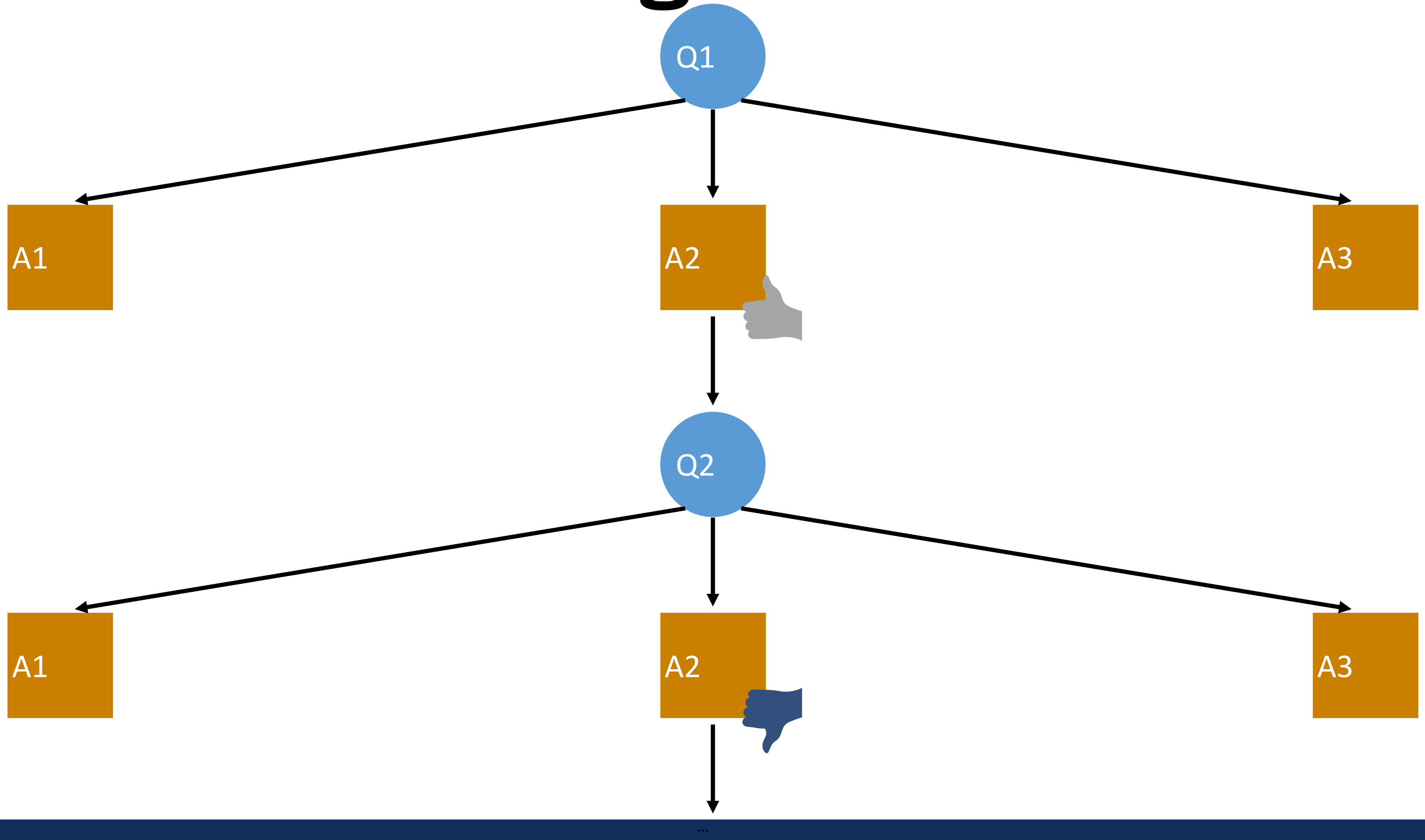
"The quick brown fox jumps over the ______"

Lazy 95%
Slow 2%
Fun 1%
…
Zyzzyva 0%

- Trained on all available text on the internet



Chat is a branching tree



Major issues with large language models

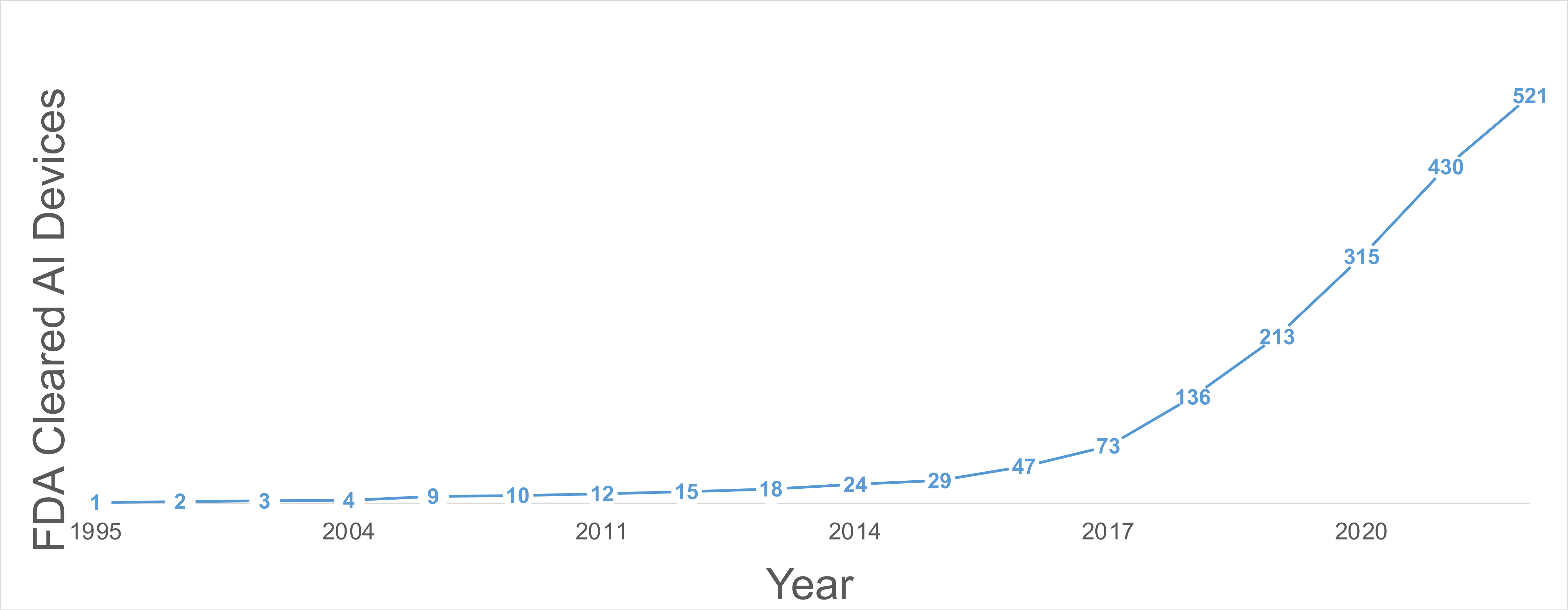
- Based on what ever data it was trained on
  - May not be relevant, accurate, or pleasant

- Generative process is inherently stochastic
  - Response choices and sentence construction depend on sampling distributions randomly

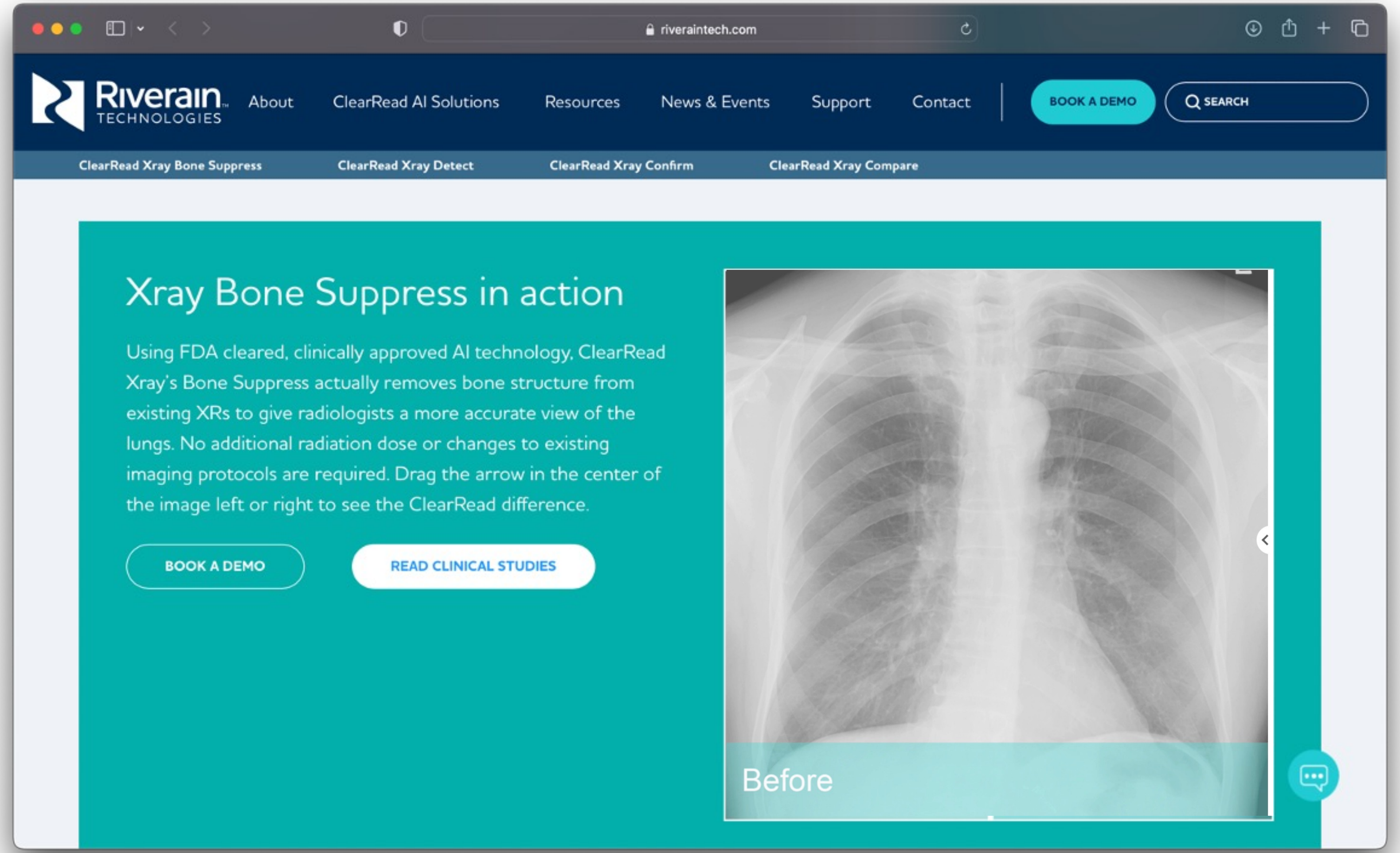
- Hard to evaluate and verify
  - How often will it be right? What is right?

How is AI used in health care?

Increasing prevalence of medical AI



AI in use at Michigan Medicine



Other examples of AI in use


THE JOURNAL of UROLOGY

Development and Validation of Models to Predict Pathological Outcomes of Radical Prostatectomy in Regional and National Cohorts

Prostate Cancer Outcomes


Mind the Performance Gap: Examining Dataset Shift During Prospective Validation

HHS Public Access

A Generalizable, Data-Driven Approach to Predict Daily Risk of *Clostridium difficile* Infection at Two Large Academic Health Centers

In Hospital Infection Risk


Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study

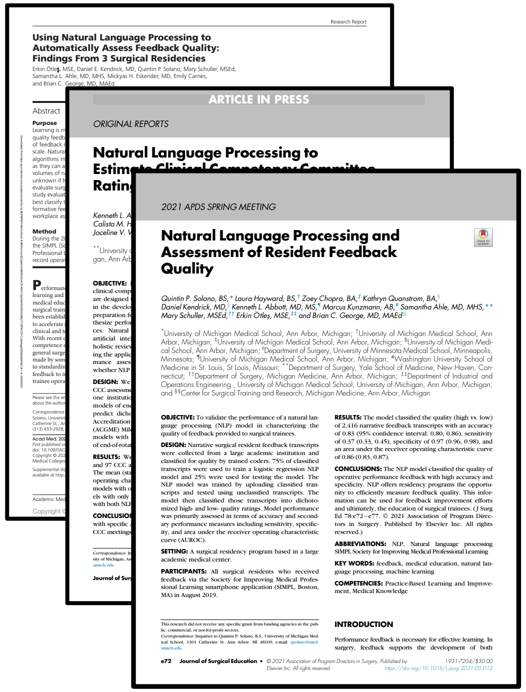
ORIGINAL RESEARCH

Evaluating a Widely Implemented Proprietary Deterioration Index Model among Hospitalized Patients with COVID-19

Deterioration Risk


External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

In Hospital Sepsis Risk


Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies

ARTICLE IN PRESS

Natural Language Processing and Assessment of Resident Feedback Quality

Feedback Evaluation

Evaluation of Proprietary Models

ORIGINAL RESEARCH

Check for updates

Evaluating a Widely Implemented Proprietary Deterioration Index Model among Hospitalized Patients with COVID-19

Karandeep Singh[1,2], Thomas S. Valley[2,3], Shengpu Tang[4], Benjamin Y. Li[4], Fahad Kamran[4], Michael W. Sjoding[2,3], Jenna Wiens[2,4], Erkin Otles[5], John P. Donnelly[1,2], Melissa Y. Wei[2,3], Jonathon P. McBride[6], Jie Cao[7], Carleen Penoza[8], John Z. Ayanian[2,3], and Brahmajee K. Nallamothu[2,3]

[1]Department of Learning Health Sciences, [3]Department of Internal Medicine, [6]Department of Cellular and Molecular Biology, and [7]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan; [2]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan; [4]Division of Computer Science and Engineering, and [5]Department of Industrial and Operations Engineering, University of Michigan College of Engineering, Ann Arbor, Michigan; and [8]Nursing Informatics, Michigan Medicine, Ann Arbor, Michigan.

ORCID IDs: 0000-0001-8980-2330 (K.S.); 0000-0002-5766-4970 (T.S.V.).

Abstract

Rationale: The Epic Deterioration Index (EDI) is a proprietary prediction model implemented in over 100 U.S. hospitals that was widely used to support medical decision-making during the coronavirus disease (COVID-19) pandemic. The EDI has not been independently evaluated, and other proprietary models have been shown to be biased against vulnerable populations.

Objectives: To independently evaluate the EDI in hospitalized patients with COVID-19 overall and in disproportionately affected subgroups.

Methods: We studied adult patients admitted with COVID-19 to units other than the intensive care unit at a large academic medical center from March 9 through May 20, 2020. We used the EDI, calculated at 15-minute intervals, to predict a composite outcome of intensive care unit–level care, mechanical ventilation, or in-hospital death. In a subset of patients hospitalized for at least 48 hours, we also evaluated the ability of the EDI to identify patients at low risk of experiencing this composite outcome during their remaining hospitalization.

Results: Among 392 COVID-19 hospitalizations meeting inclusion criteria, 103 (26%) met the composite outcome. The median age of the cohort was 64 (interquartile range, 53–75) with 168 (43%) Black patients and 169 (43%) women. The area under the receiver-operating characteristic curve of the EDI was 0.79 (95% confidence interval, 0.74–0.84). EDI predictions did not differ by race or sex. When exploring clinically relevant thresholds of the EDI, we found patients who met or exceeded an EDI of 68.8 made up 14% of the study cohort and had a 74% probability of experiencing the composite outcome during their hospitalization with a sensitivity of 39% and a median lead time of 24 hours from when this threshold was first exceeded. Among the 286 patients hospitalized for at least 48 hours who had not experienced the composite outcome, 14 (13%) never exceeded an EDI of 37.9, with a negative predictive value of 90% and a sensitivity above this threshold of 91%.

Conclusions: We found the EDI identifies small subsets of high-risk and low-risk patients with COVID-19 with good discrimination, although its clinical use as an early warning system is limited by low sensitivity. These findings highlight the importance of independent evaluation of proprietary models before widespread operational use among patients with COVID-19.

Keywords: coronavirus disease; deterioration index; prediction model; validation study

(Received in original form June 19, 2020; accepted in final form December 23, 2020)

Supported by National Heart, Lung, and Blood Institute grants K23 140165 (T.S.V.), K01 HL136687 (M.W.S.), and K12-HL138039 (J.P.D.); National Institute of Aging grant K23 AG056638 (M.Y.W.); National Library of Medicine and the Michigan Institute for Data Science grant R01 LM013325-01 (M.W.S., J.W., and B.K.N.); National Institute of General Medical Sciences grant T32GM007863 (B.Y.L.); and U.S. National Institutes of Health grant T32GM007863 (E.O.). J.W. served on the board of the nonprofit organization Machine Learning for Healthcare.

Author Contributions: K.S., T.S.V., M.W.S., J.W., J.P.D., M.Y.W., J.P.M., J.C., C.P., J.Z.A., and B.K.N. made substantial contributions to the conception or design of the work. K.S., S.T., B.Y.L., F.K., M.W.S., J.W., and E.O. participated in the data acquisition and analysis. All authors participated in the interpretation of data. K.S. drafted the manuscript and all authors participated in critical revisions for intellectual content. All authors gave approval for the version being considered for publication and agreed to be accountable for the analysis, findings, and interpretation.

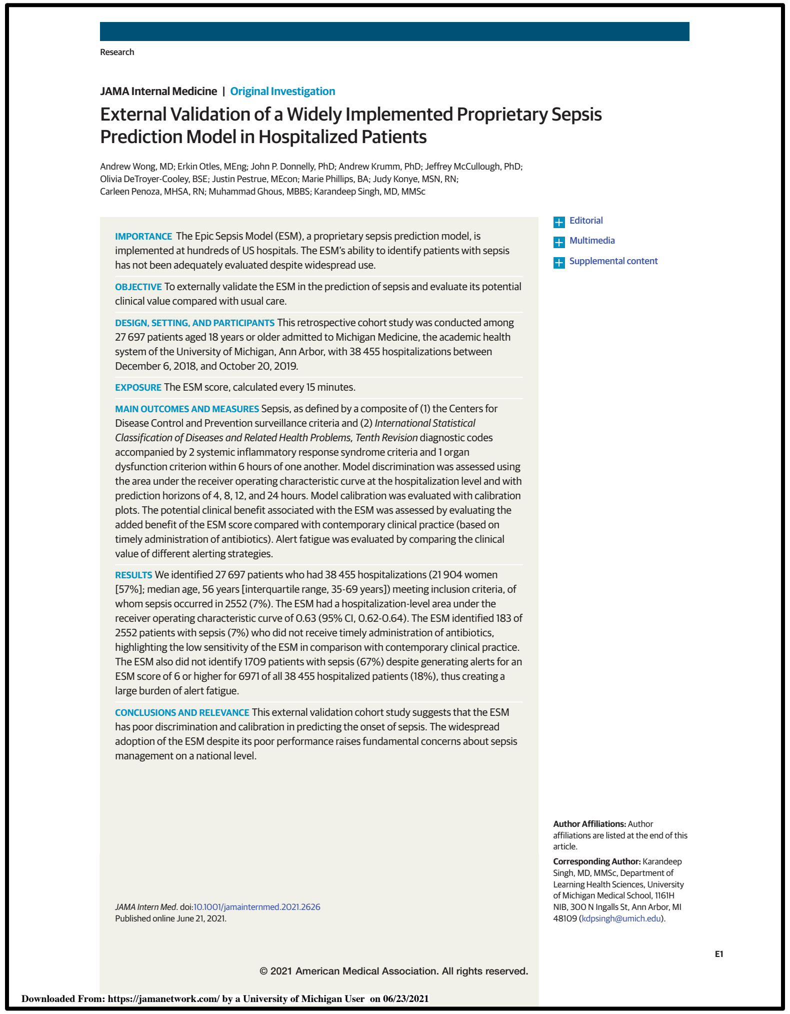
Correspondence and requests for reprints should be addressed to Karandeep Singh, M.D., M. M.Sc., Department of Learning Health Sciences, University of Michigan Medical School, 1161H NIB, 300 N. Ingalls Street, Ann Arbor, MI 48109. E-mail: kdpsingh@umich.edu.

This article has a related editorial.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Ann Am Thorac Soc Vol 18, No 7, pp 1129–1137, Jul 2021

DOI: 10.1513/AnnalsATS.202006-698OC
Internet address: www.atsjournals.org



Research

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

Editorial

Multimedia

Supplemental content

IMPORTANCE The Epic Sepsis Model (ESM), a proprietary sepsis prediction model, is implemented at hundreds of US hospitals. The ESM's ability to identify patients with sepsis has not been adequately evaluated despite widespread use.

OBJECTIVE To externally validate the ESM in the prediction of sepsis and evaluate its potential clinical value compared with usual care.

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study was conducted among 27 697 patients aged 18 years or older admitted to Michigan Medicine, the academic health system of the University of Michigan, Ann Arbor, with 38 455 hospitalizations between December 6, 2018, and October 20, 2019.

EXPOSURE The ESM score, calculated every 15 minutes.

MAIN OUTCOMES AND MEASURES Sepsis, as defined by a composite of (1) the Centers for Disease Control and Prevention surveillance criteria and (2) *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* diagnostic codes accompanied by 2 systemic inflammatory response syndrome criteria and 1 organ dysfunction criterion within 6 hours of one another. Model discrimination was assessed using the area under the receiver operating characteristic curve at the hospitalization level and with prediction horizons of 4, 8, 12, and 24 hours. Model calibration was evaluated with calibration plots. The potential clinical benefit associated with the ESM was assessed by evaluating the added benefit of the ESM score compared with contemporary clinical practice (based on timely administration of antibiotics). Alert fatigue was evaluated by comparing the clinical value of different alerting strategies.

RESULTS We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

CONCLUSIONS AND RELEVANCE This external validation cohort study suggests that the ESM has poor discrimination and calibration in predicting the onset of sepsis. The widespread adoption of the ESM despite its poor performance raises fundamental concerns about sepsis management on a national level.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Karandeep Singh, MD, MMSc, Department of Learning Health Sciences, University of Michigan Medical School, 1161H NIB, 300 N Ingalls St, Ann Arbor, MI 48109 (kdpsingh@umich.edu).

JAMA Intern Med. doi:10.1001/jamainternmed.2021.2626
Published online June 21, 2021.

Singh 2021, Wong 2021

UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
MICHIGAN MEDICINE

Use of ML in Medical Trainee Feedback

Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies

Erkin Ötleş, MSE, Daniel E. Kendrick, MD, Quintin P. Solano, Mary Schuller, MSEd, Samantha L. Ahle, MD, MHS, Mickyas H. Eskender, MD, Emily Carnes, and Brian C. George, MD, MAEd

Abstract

Purpose
Learning is markedly improved with high-quality feedback, yet assuring the quality of feedback is difficult to achieve at scale. Natural language processing (NLP) algorithms may be useful in this context as they can automatically classify large volumes of narrative data. However, it is unknown if NLP models can accurately evaluate surgical trainee feedback. This study evaluated which NLP techniques best classify the quality of surgical trainee formative feedback recorded as part of a workplace assessment.

Method
During the 2016–2017 academic year, the SIMPL (Society for Improving Medical Professional Learning) app was used to record operative performance narrative feedback for residents at 3 university-based general surgery residency training programs. Feedback comments were collected for a sample of residents representing all 5 postgraduate year levels and coded for quality. In May 2019, the coded comments were then used to train NLP models to automatically classify the quality of feedback across 4 categories (effective, mediocre, ineffective, or other). Models included support vector machines (SVM), logistic regression, gradient boosted trees, naive Bayes, and random forests. The primary outcome was mean classification accuracy.

Results
The authors manually coded the quality of 600 recorded feedback comments. Those data were used to train NLP models to automatically classify the quality of feedback across 4 categories. The NLP model using an SVM algorithm yielded a maximum mean accuracy of 0.64 (standard deviation, 0.01). When the classification task was modified to distinguish only high-quality vs low-quality feedback, maximum mean accuracy was 0.83, again with SVM.

Conclusions
To the authors' knowledge, this is the first study to examine the use of NLP for classifying feedback quality. SVM NLP models demonstrated the ability to automatically classify the quality of surgical trainee evaluations. Larger training datasets would likely further increase accuracy.

Performance feedback is critical to learning and is highly valued across medical education domains.[1–3] In surgical training, its significance has been established as a powerful means to accelerate improvement in both clinical and technical performance.[4–7] With recent concerns regarding the competence of graduating residents in general surgery,[8,9] an effort has been made by some surgical training programs to standardize the components of quality feedback to improve the assessment of trainee operative performance.[10,11]

Please see the end of this article for information about the authors.

Correspondence should be addressed to Quintin P. Solano, University of Michigan Medical School, 1301 Catherine St., Ann Arbor, MI 48109; telephone: (313) 433-2928; email: qsolano@med.umich.edu.

Acad Med. 2021;96:1457–1460.
First published online May 4, 2021
doi: 10.1097/ACM.0000000000004153

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/B110.

Unfortunately, it is difficult to ensure that faculty are adhering to these standards when delivering feedback to residents in practice. Current methods to evaluate feedback quality are labor intensive because they require trained raters to review and classify the quality of individual recorded feedback.[12,13] This is a growing problem with the widespread adoption of smartphone assessment applications, which have greatly increased the volume of narrative feedback available to trainees.[14–16] Alternative methodologies for feedback quality assurance are therefore needed to efficiently identify instructors who are not meeting standards and who might benefit most from targeted faculty development.

Emerging technology in machine learning (ML) may be an automated solution. A subfield of ML, known as natural language processing (NLP), includes algorithms developed for automated text analysis. NLP has been successfully developed in other fields to classify document sentiment, identify important entities in text, and even automatically translate text from one language to another. In medical education, a variety of NLP techniques have been used to automate the evaluation of trainee documentation and clinical experiences.[17–19] However, to our knowledge, NLP techniques have never been used to assess the quality of feedback provided to trainees by faculty.[20–22] In an effort to better understand how automated feedback quality assurance could be implemented using NLP, we investigated the accuracy of different NLP models to classify the quality of feedback provided to surgical trainees.

Method

Study population

We conducted this analysis in May 2019 at the University of Michigan Medical School. Data were collected from a convenience sample of 3 university-based general surgery residency training programs, all part of large

---

Natural Language Processing to Estimate Clinical Competency Committee Ratings

Kenneth L. Abbott, MD, MS,** Brian C. George, MD, MAEd,† Gurjit Sandhu, PhD,† Calista M. Harbaugh, MD, MS,† Paul G. Gauger, MD,† Erkin Ötleş, MEng,** Niki Matusko, BS,† and Joceline V. Vu, MD†

**University of Michigan Medical School, Ann Arbor, Michigan; and †Department of Surgery, University of Michigan, Ann Arbor, Michigan

OBJECTIVE: Residency program faculty participate in clinical competency committee (CCC) meetings, which are designed to evaluate residents' performance and aid in the development of individualized learning plans. In preparation for the CCC meetings, faculty members synthesize performance information from a variety of sources. Natural language processing (NLP), a form of artificial intelligence, might facilitate these complex holistic reviews. However, there is little research involving the application of this technology to resident performance assessments. With this study, we examine whether NLP can be used to estimate CCC ratings.

DESIGN: We analyzed end-of-rotation assessments and CCC assessments for all surgical residents who trained at one institution between 2014 and 2018. We created models of end-of-rotation assessment ratings and text to predict dichotomized CCC assessment ratings for 16 Accreditation Council for Graduate Medical Education (ACGME) Milestones. We compared the performance of models with and without predictors derived from NLP of end-of-rotation assessment text.

RESULTS: We analyzed 594 end-of-rotation assessments and 97 CCC assessments for 24 general surgery residents. The mean (standard deviation) for area under the receiver operating characteristic curve (AUC) was 0.84 (0.05) for models with only non-NLP predictors, 0.83 (0.06) for models with only NLP predictors, and 0.87 (0.05) for models with both NLP and non-NLP predictors.

CONCLUSIONS: NLP can identify language correlated with specific ACGME Milestone ratings. In preparation for CCC meetings, faculty could use information automatically extracted from text to focus attention on residents who might benefit from additional support and guide the development of educational interventions. (J Surg Ed 000:1–6. © 2021 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Natural language processing, clinical competency committee, resident, assessment, evaluation

COMPETENCIES: Patient Care, Medical Knowledge, Systems-Based Practice, Practice-Based Learning And Improvement, Professionalism, Interpersonal And Communication Skills

INTRODUCTION

Residency programs use a system of assessments to track trainee progress and development. For example, a subset of faculty members participates in clinical competency committee (CCC) meetings, which occur every six months and are designed to evaluate performance and aid in the development of individualized learning plans and interventions.[1] In preparation for the CCC meetings, committee members synthesize performance information from a variety of sources—some formal (e.g., monthly end-of-rotation assessments) and some informal (e.g., conversations).

Artificial intelligence could support the CCC faculty performing these complex holistic reviews by guiding their attention to residents who may benefit from additional support. Natural language processing (NLP) is a form of artificial intelligence that interprets complex human language.[2] In general surgery, Milestones are used to structure CCC meeting discussion and resident assessment.[3,4] It is unknown whether NLP can identify language correlated with specific Accreditation Council

Correspondence: Inquiries to Joceline V. Vu MD, Department of Surgery, University of Michigan, Ann Arbor, MI; 48109. Phone: (703) 200-8623; e-mail: vuj@med.umich.edu

---

Natural Language Processing and Assessment of Resident Feedback Quality

Quintin P. Solano, BS,* Laura Hayward, BS,† Zoey Chopra, BA,‡ Kathryn Quanstrom, BA,§ Daniel Kendrick, MD,|| Kenneth L. Abbott, MD, MS,¶ Marcus Kunzmann, AB,# Samantha Ahle, MD, MHS,** Mary Schuller, MSEd,†† Erkin Ötleş, MSE,‡‡ and Brian C. George, MD, MAEd§§

*University of Michigan Medical School, Ann Arbor, Michigan; †University of Michigan Medical School, Ann Arbor, Michigan; ‡University of Michigan Medical School, Ann Arbor, Michigan; §University of Michigan Medical School, Ann Arbor, Michigan; ||Department of Surgery, University of Minnesota Medical School, Minneapolis, Minnesota; ¶University of Michigan Medical School, Ann Arbor, Michigan; #Washington University School of Medicine in St. Louis, St Louis, Missouri; **Department of Surgery, Yale School of Medicine, New Haven, Connecticut; ††Department of Surgery, Michigan Medicine, Ann Arbor, Michigan; ‡‡Department of Industrial and Operations Engineering , University of Michigan Medical School, University of Michigan, Ann Arbor, Michigan; and §§Center for Surgical Training and Research, Michigan Medicine, Ann Arbor, Michigan

OBJECTIVE: To validate the performance of a natural language processing (NLP) model in characterizing the quality of feedback provided to surgical trainees.

DESIGN: Narrative surgical resident feedback transcripts were collected from a large academic institution and classified for quality by trained coders. 75% of classified transcripts were used to train a logistic regression NLP model and 25% were used for testing the model. The NLP model was trained by uploading classified transcripts and tested using unclassified transcripts. The model then classified those transcripts into dichotomized high- and low- quality ratings. Model performance was primarily assessed in terms of accuracy and secondary performance measures including sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC).

SETTING: A surgical residency program based in a large academic medical center.

PARTICIPANTS: All surgical residents who received feedback via the Society for Improving Medical Professional Learning smartphone application (SIMPL, Boston, MA) in August 2019.

RESULTS: The model classified the quality (high vs. low) of 2,416 narrative feedback transcripts with an accuracy of 0.83 (95% confidence interval: 0.80, 0.86), sensitivity of 0.37 (0.33, 0.45), specificity of 0.97 (0.96, 0.98), and an area under the receiver operating characteristic curve of 0.86 (0.83, 0.87).

CONCLUSIONS: The NLP model classified the quality of operative performance feedback with high accuracy and specificity. NLP offers residency programs the opportunity to efficiently measure feedback quality. This information can be used for feedback improvement efforts and ultimately, the education of surgical trainees. (J Surg Ed 78:e72–e77. © 2021 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

ABBREVIATIONS: NLP, Natural language processing SIMPL Society for Improving Medical Professional Learning

KEY WORDS: feedback, medical education, natural language processing, machine learning

COMPETENCIES: Practice-Based Learning and Improvement, Medical Knowledge

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Correspondence: Inquiries to Quintin P. Solano, B.S., University of Michigan Medical School, 1301 Catherine St. Ann Arbor, MI 48109; e-mail: qsolano@med.umich.edu

INTRODUCTION

Performance feedback is necessary for effective learning. In surgery, feedback supports the development of both

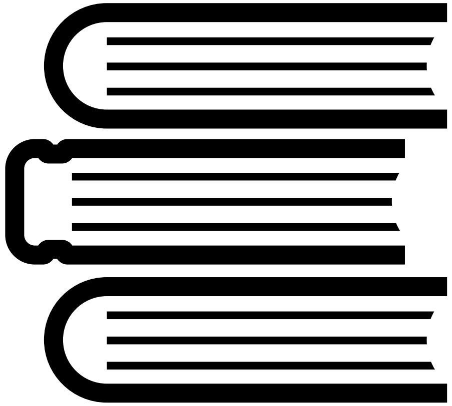
---

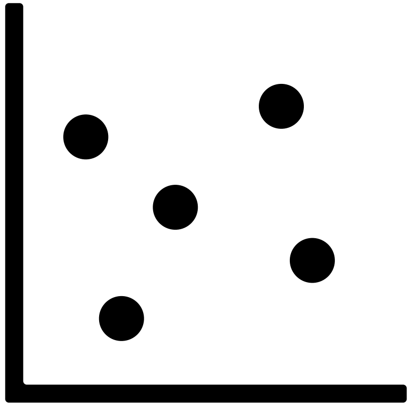
Ötleş 2021, Abbott 2021, Solano 2021

UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
MICHIGAN MEDICINE

Why should we train physicians on AI?

AI has the potential to advance medicine



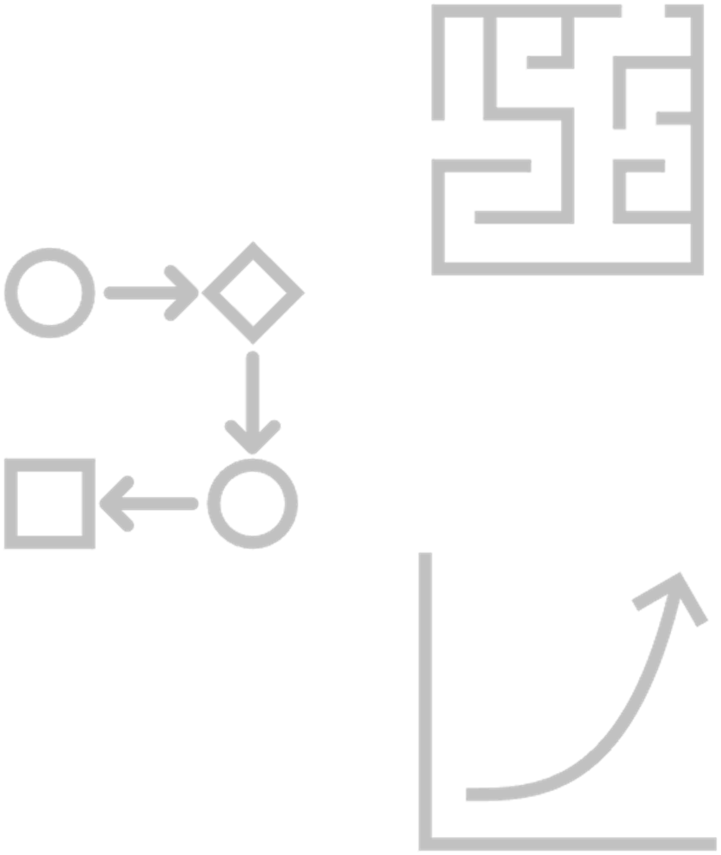


- AI has techniques to rapidly **summarize** information, **predict** outcomes, and **learn** over time

- Society has big expectations for AI in medicine

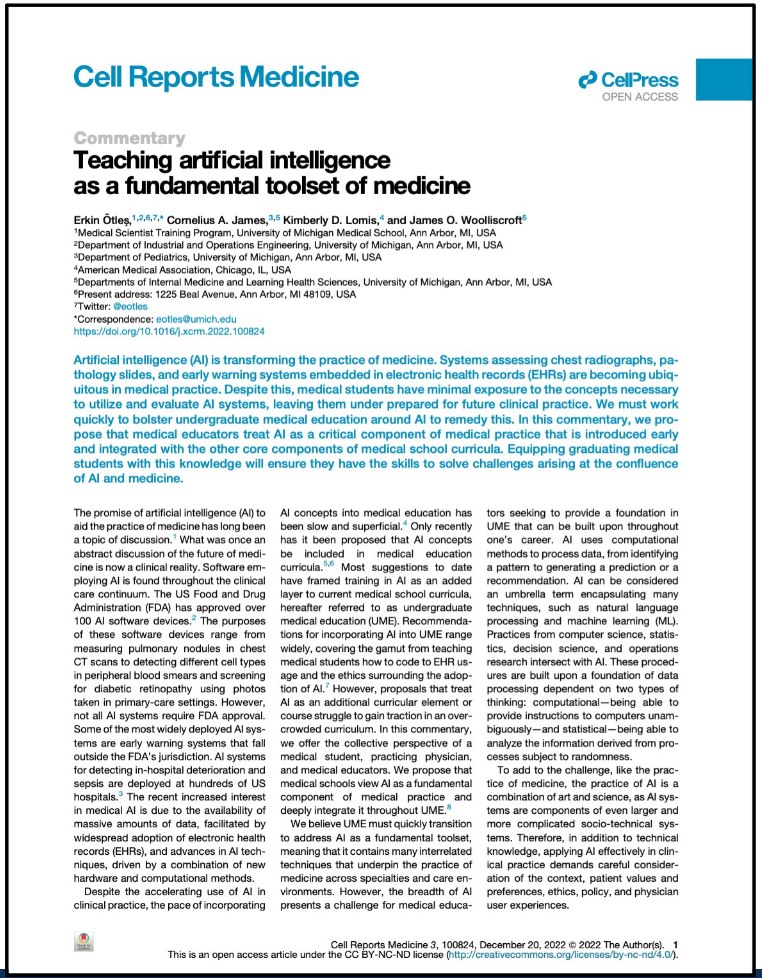
AI is not a part of medical education

- Use of AI in medicine is not straightforward
- AI tools depend on complicated data and workflows that physicians understand
- Medical AI adoption increasing
- Learners unprepared to use, assess, and develop AI tools



We’ve got to start training physicians on AI fundamentals

- Physicians shouldn’t just be “users”
- Should be actively involved in creating, evaluating, and improving AI
- Leadership in AI dependent on:
  - **understanding** how it works
  - **partnership** with engineers



Cell Reports Medicine

CellPress
OPEN ACCESS

Commentary

Teaching artificial intelligence as a fundamental toolset of medicine

Erkin Ötleş,[1,2,6,7,*] Cornelius A. James,[3,5] Kimberly D. Lomis,[4] and James O. Woolliscroft[5]
[1]Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, MI, USA
[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, USA
[3]Department of Pediatrics, University of Michigan, Ann Arbor, MI, USA
[4]American Medical Association, Chicago, IL, USA
[5]Departments of Internal Medicine and Learning Health Sciences, University of Michigan, Ann Arbor, MI, USA
[6]Present address: 1225 Beal Avenue, Ann Arbor, MI 48109, USA
[7]Twitter: @eotles
*Correspondence: eotles@umich.edu
https://doi.org/10.1016/j.xcrm.2022.100824

Artificial intelligence (AI) is transforming the practice of medicine. Systems assessing chest radiographs, pathology slides, and early warning systems embedded in electronic health records (EHRs) are becoming ubiquitous in medical practice. Despite this, medical students have minimal exposure to the concepts necessary to utilize and evaluate AI systems, leaving them under prepared for future clinical practice. We must work quickly to bolster undergraduate medical education around AI to remedy this. In this commentary, we propose that medical educators treat AI as a critical component of medical practice that is introduced early and integrated with the other core components of medical school curricula. Equipping graduating medical students with this knowledge will ensure they have the skills to solve challenges arising at the confluence of AI and medicine.

The promise of artificial intelligence (AI) to aid the practice of medicine has long been a topic of discussion.[1] What was once an abstract discussion of the future of medicine is now a clinical reality. Software employing AI is found throughout the clinical care continuum. The US Food and Drug Administration (FDA) has approved over 100 AI software devices.[2] The purposes of these software devices range from measuring pulmonary nodules in chest CT scans to detecting different cell types in peripheral blood smears and screening for diabetic retinopathy using photos taken in primary-care settings. However, not all AI systems require FDA approval. Some of the most widely deployed AI systems are early warning systems that fall outside the FDA’s jurisdiction. AI systems for detecting in-hospital deterioration and sepsis are deployed at hundreds of US hospitals.[3] The recent increased interest in medical AI is due to the availability of massive amounts of data, facilitated by widespread adoption of electronic health records (EHRs), and advances in AI techniques, driven by a combination of new hardware and computational methods.

Despite the accelerating use of AI in clinical practice, the pace of incorporating AI concepts into medical education has been slow and superficial.[4] Only recently has it been proposed that AI concepts be included in medical education curricula.[5,6] Most suggestions to date have framed training in AI as an added layer to current medical school curricula, hereafter referred to as undergraduate medical education (UME). Recommendations for incorporating AI into UME range widely, covering the gamut from teaching medical students how to code to EHR usage and the ethics surrounding the adoption of AI.[7] However, proposals that treat AI as an additional curricular element or course struggle to gain traction in an overcrowded curriculum. In this commentary, we offer the collective perspective of a medical student, practicing physician, and medical educators. We propose that medical schools view AI as a fundamental component of medical practice and deeply integrate it throughout UME.[8]

We believe UME must quickly transition to address AI as a fundamental toolset, meaning that it contains many interrelated techniques that underpin the practice of medicine across specialties and care environments. However, the breadth of AI presents a challenge for medical educators seeking to provide a foundation in UME that can be built upon throughout one’s career. AI uses computational methods to process data, from identifying a pattern to generating a prediction or a recommendation. AI can be considered an umbrella term encapsulating many techniques, such as natural language processing and machine learning (ML). Practices from computer science, statistics, decision science, and operations research intersect with AI. These procedures are built upon a foundation of data processing dependent on two types of thinking: computational—being able to provide instructions to computers unambiguously—and statistical—being able to analyze the information derived from processes subject to randomness.

To add to the challenge, like the practice of medicine, the practice of AI is a combination of art and science, as AI systems are components of even larger and more complicated socio-technical systems. Therefore, in addition to technical knowledge, applying AI effectively in clinical practice demands careful consideration of the context, patient values and preferences, ethics, policy, and physician user experiences.

Are you currently using AI for teaching (instruction, assessment)?

Are you currently teaching about the role of AI in health care?


EDUCATION

Opinion

VIEWPOINT

Artificial Intelligence in Health Care
A Report From the National Academy of Medicine

- Promote population-representative data with accessibility, standardization and quality is imperative.
- Prioritize ethical, equitable and inclusive medical AI while addressing explicit and implicit bias.
- Contextualize the dialogue of transparency and trust, which means accepting differential needs.
- Focus in the near term on augmented intelligence rather than autonomous agents.
- **Develop and deploy appropriate training and educational programs.**
- Leverage frameworks and best practices for learning health care systems, human factors and implementation science.
- Balance innovation with safety through regulation and legislation to promote trust.

DISCUSSION PAPER

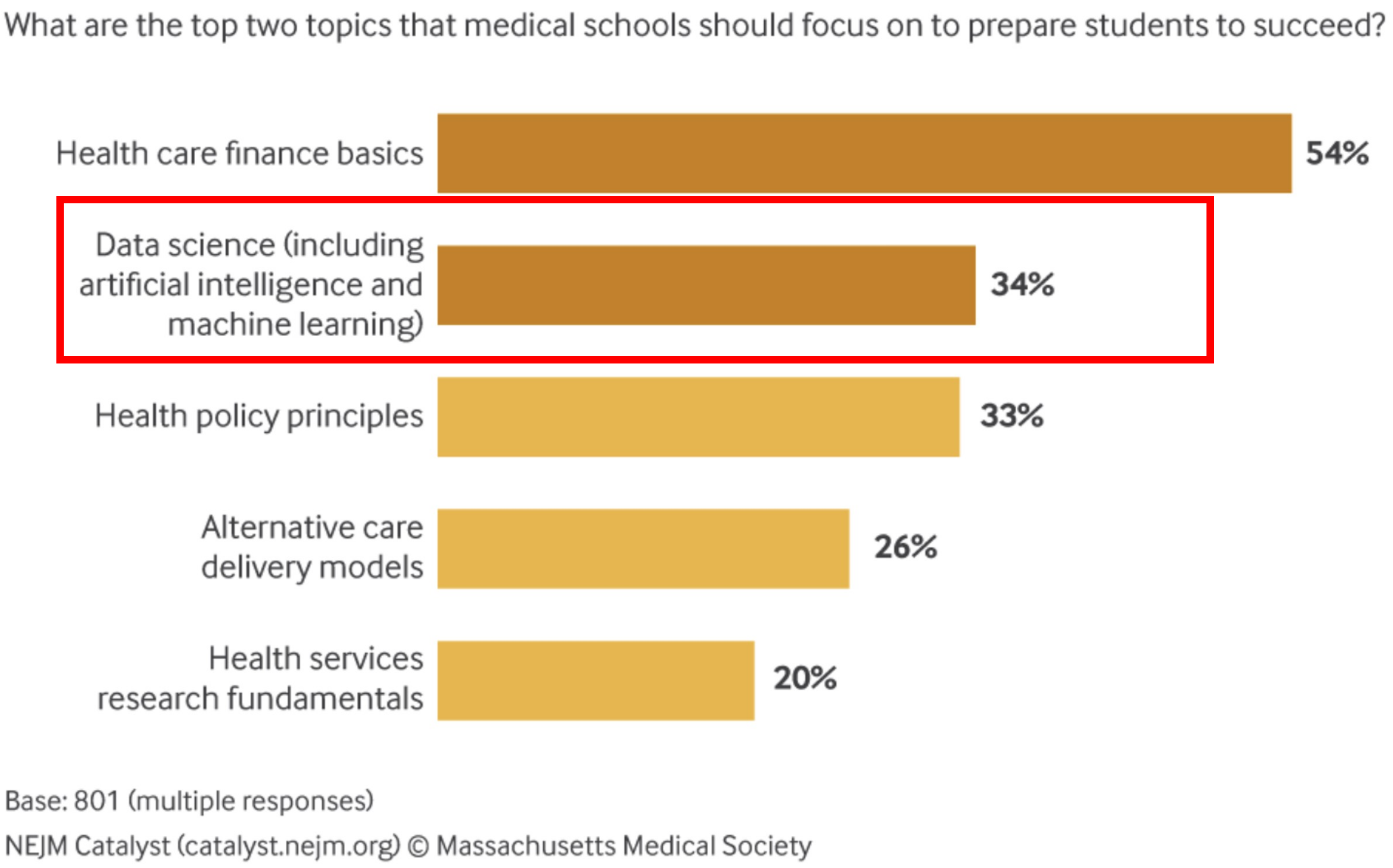
Artificial Intelligence for Health Professions Educators

Kimberly Lomis, MD, American Medical Association; **Pamela Jeffries, PHD, RN, FAAN, ANEF,** Vanderbilt School of Nursing; **Anthony Palatta, DDS, EdD,** PalattaSolutions; **Melanie Sage, PHD, MSW,** University at Buffalo School of Social Work; **Javaid Sheikh, MD, MBA,** Weill Cornell Medicine-Qatar; **Carl Sheperis, PhD, MS,** Texas A&M University-San Antonio; and **Alison Whelan, MD,** Association of American Medical Colleges

September 8, 2021

James CA, Wachter RM, Woolliscroft JO. Preparing Clinicians for a Clinical World Influenced by Artificial Intelligence. *JAMA*. 2022;327(14):1333-1334.

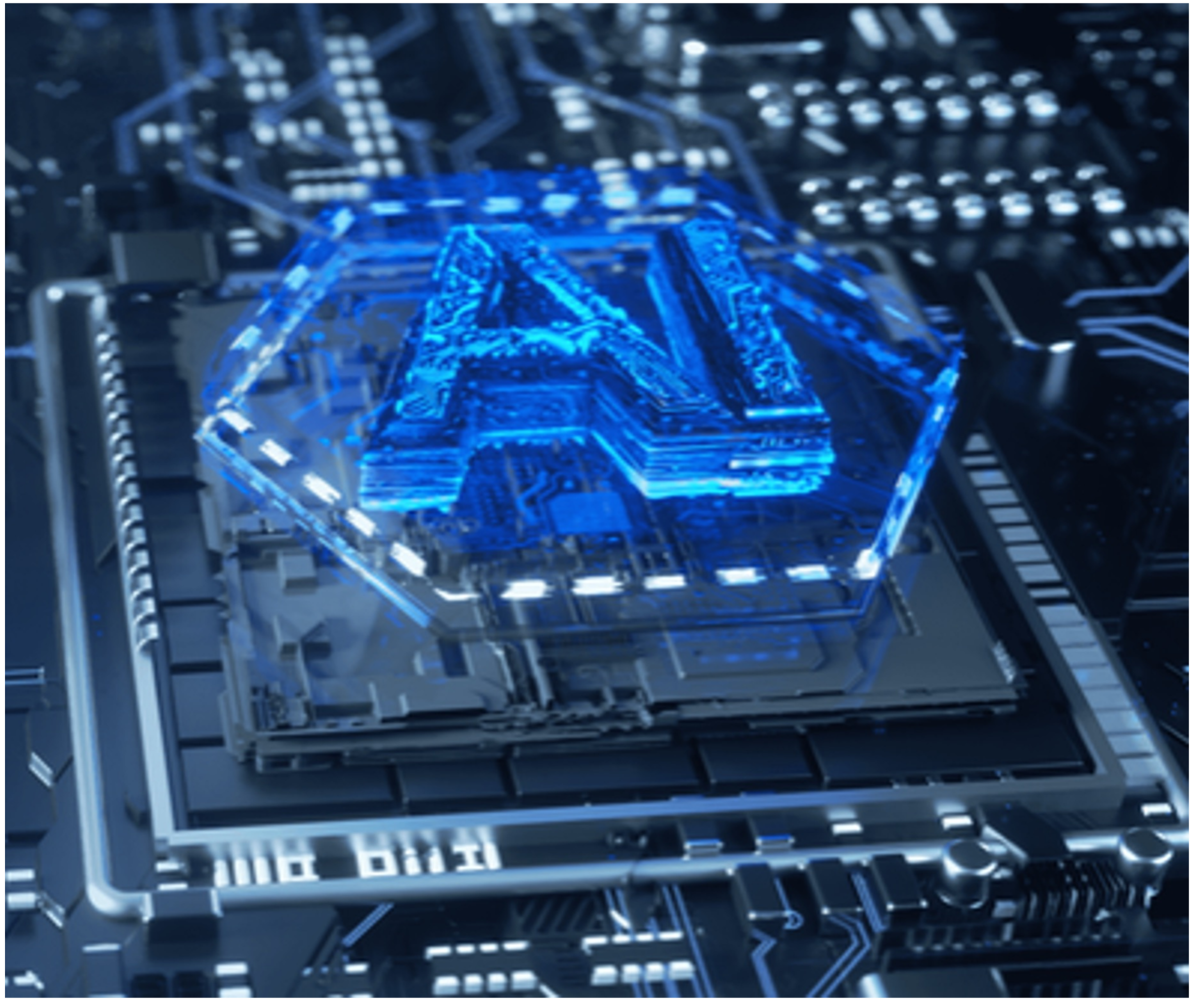
NEJM Poll



Mohta N, Johnston SC. Medical education in need of a 2020 revamp. *NEJM Catalyst.* 2020;1(3):1-7.

Current State

- Electives
- Online courses, modules
- Workshops
- Certificate programs
- Interest groups



1. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing Artificial Intelligence Training in Medical Education. *JMIR Med Educ*. 2019;5(2):e16048.
2. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. *Acad Med*. 2021;96(11S):S62-S70.

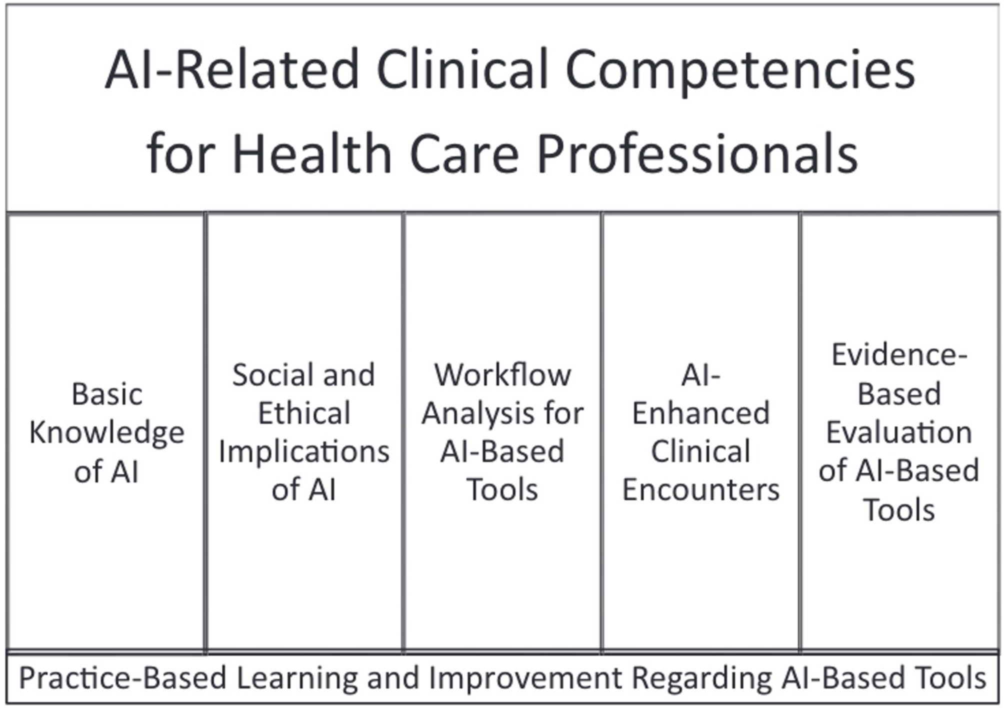
Goals of AI/ML Instruction

- Data-savvy consumers
- Patient advocacy
- Fundamental concepts
- Appraisal, evaluation
- Clinical application
- Biases, legal, ethical considerations
  - Clinical and systems level
- Data stewardship and data quality assurance

Shift focus from “information acquisition” to “information management”

Competencies for the Use of Artificial Intelligence–Based Tools by Health Care Professionals

Regina G. Russell, PhD, MA, MEd, Laurie Lovett Novak, PhD, Mehool Patel, MD, Kim V. Garvey, PhD, MS, MLIS, Kelly Jean Thomas Craig, PhD, Gretchen P. Jackson, MD, PhD, Don Moore, PhD, and Bonnie M. Miller, MD, MMHC



JMIR MEDICAL EDUCATION Weidener & Fischer

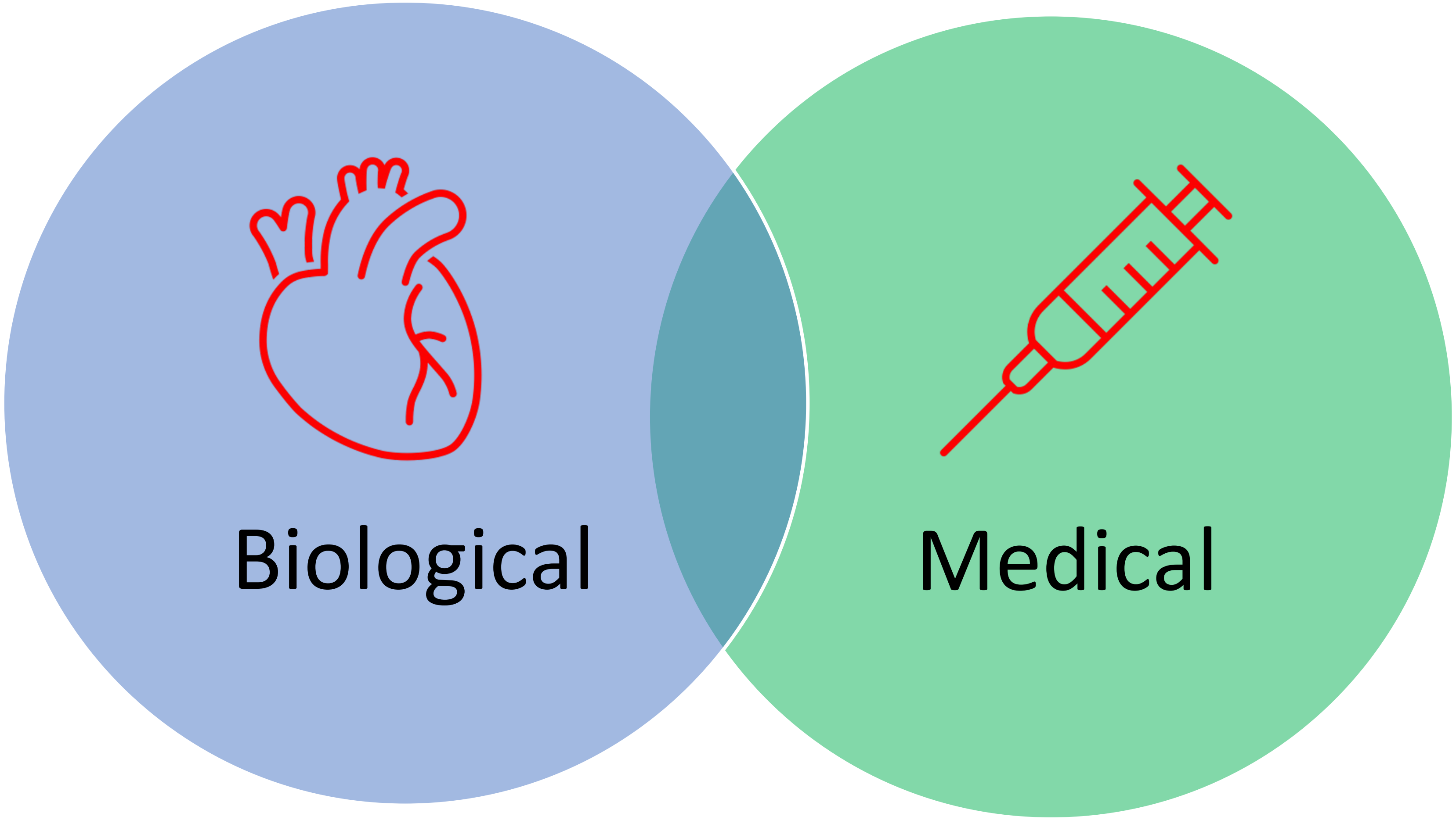
Original Paper

Artificial Intelligence Teaching as Part of Medical Education: Qualitative Analysis of Expert Interviews

Table 1. Overview of the 3 defined main categories with the associated 9 subcategories.

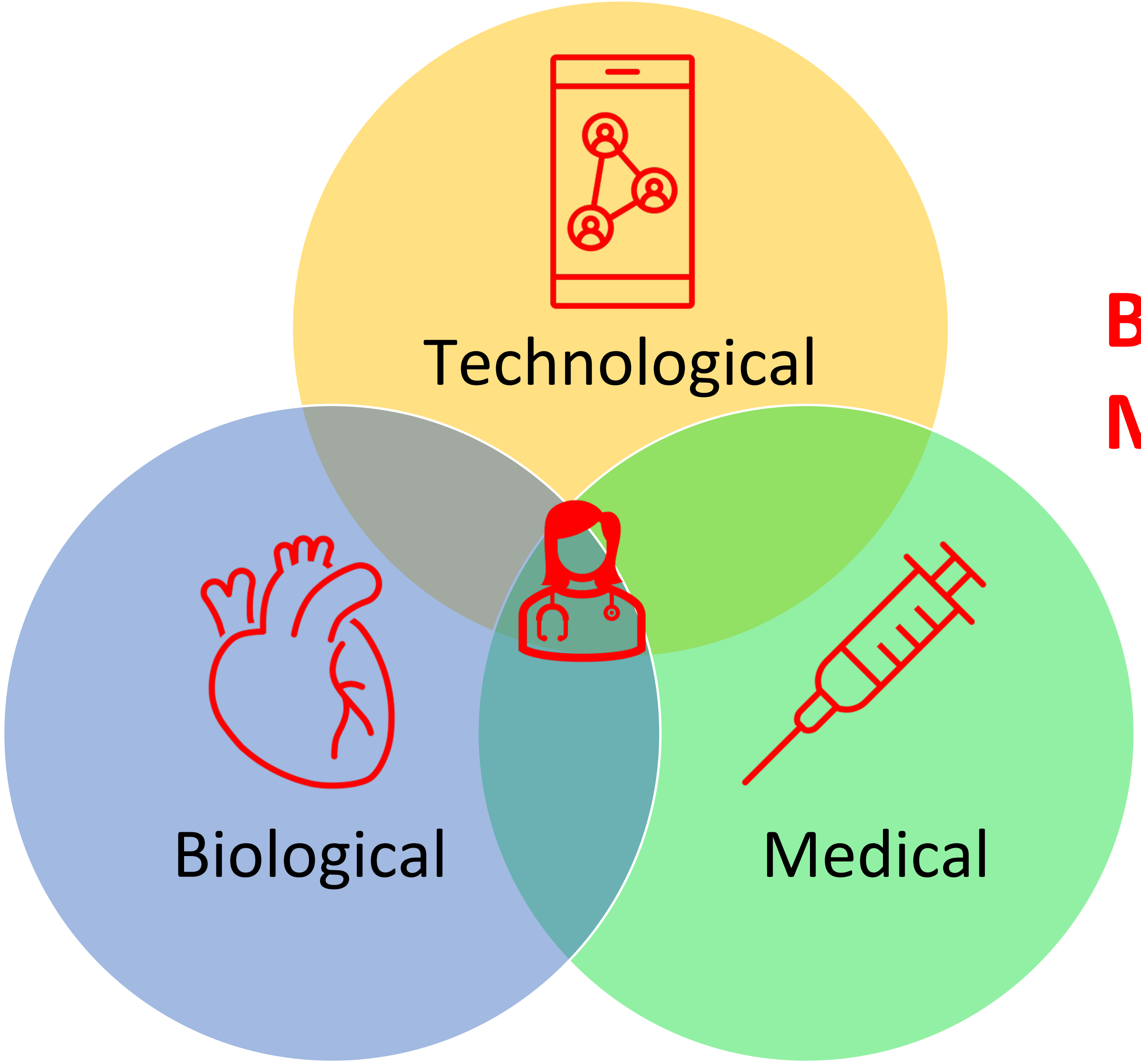
Main categories	Subcategories
Knowledge	• Basic understanding of artificial intelligence • Statistics • Ethics • Data protection and regulation
Interpretation	• Critical reflection • Associated risks • Data basis
Application	• Practical skills • Trust

McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? *NPJ Digital Medicine*. 2020;3:86.

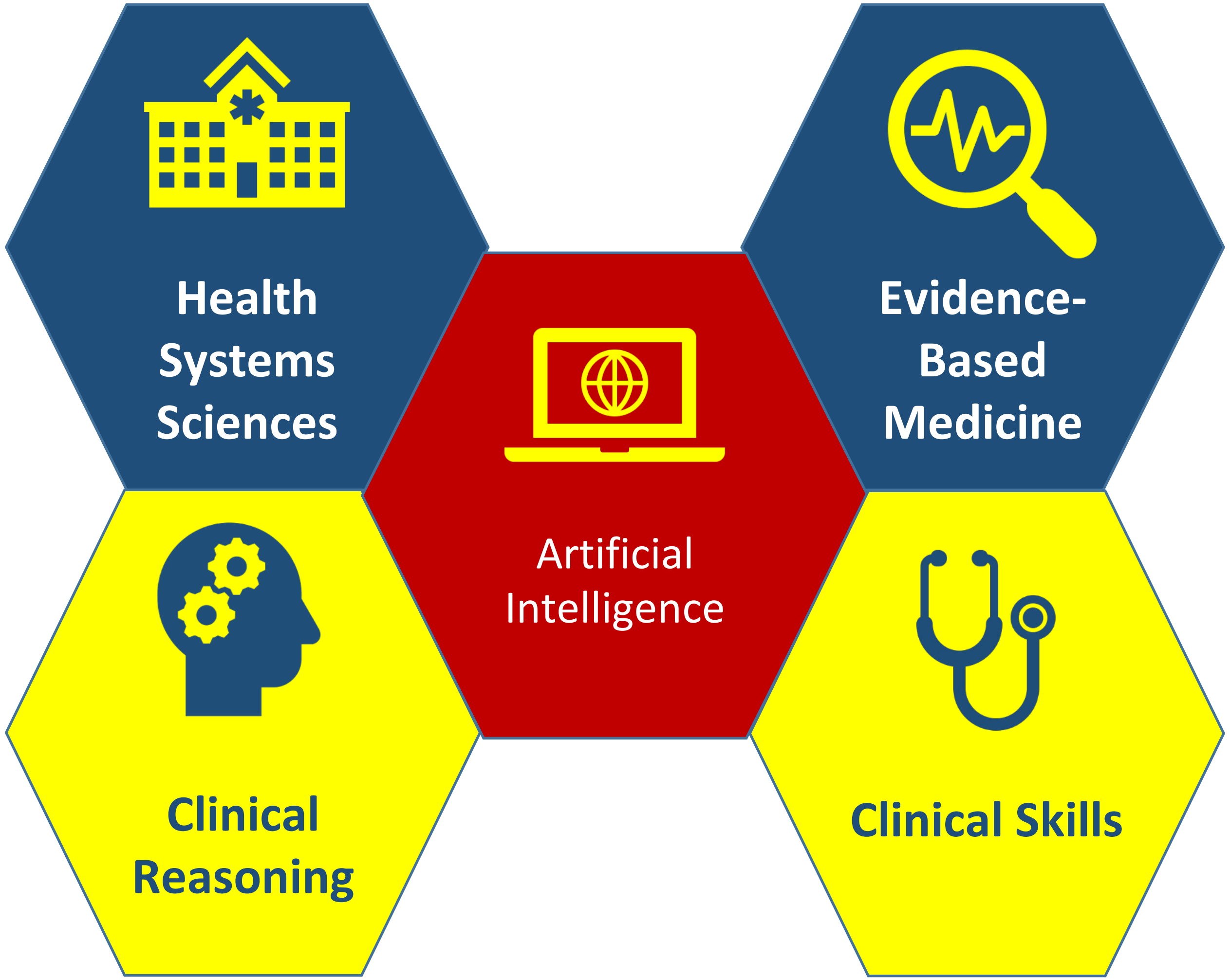


Biomedical Model

Duffy TP. The Flexner Report--100 years later. *Yale J Biol Med*. 2011;84(3):269-276.



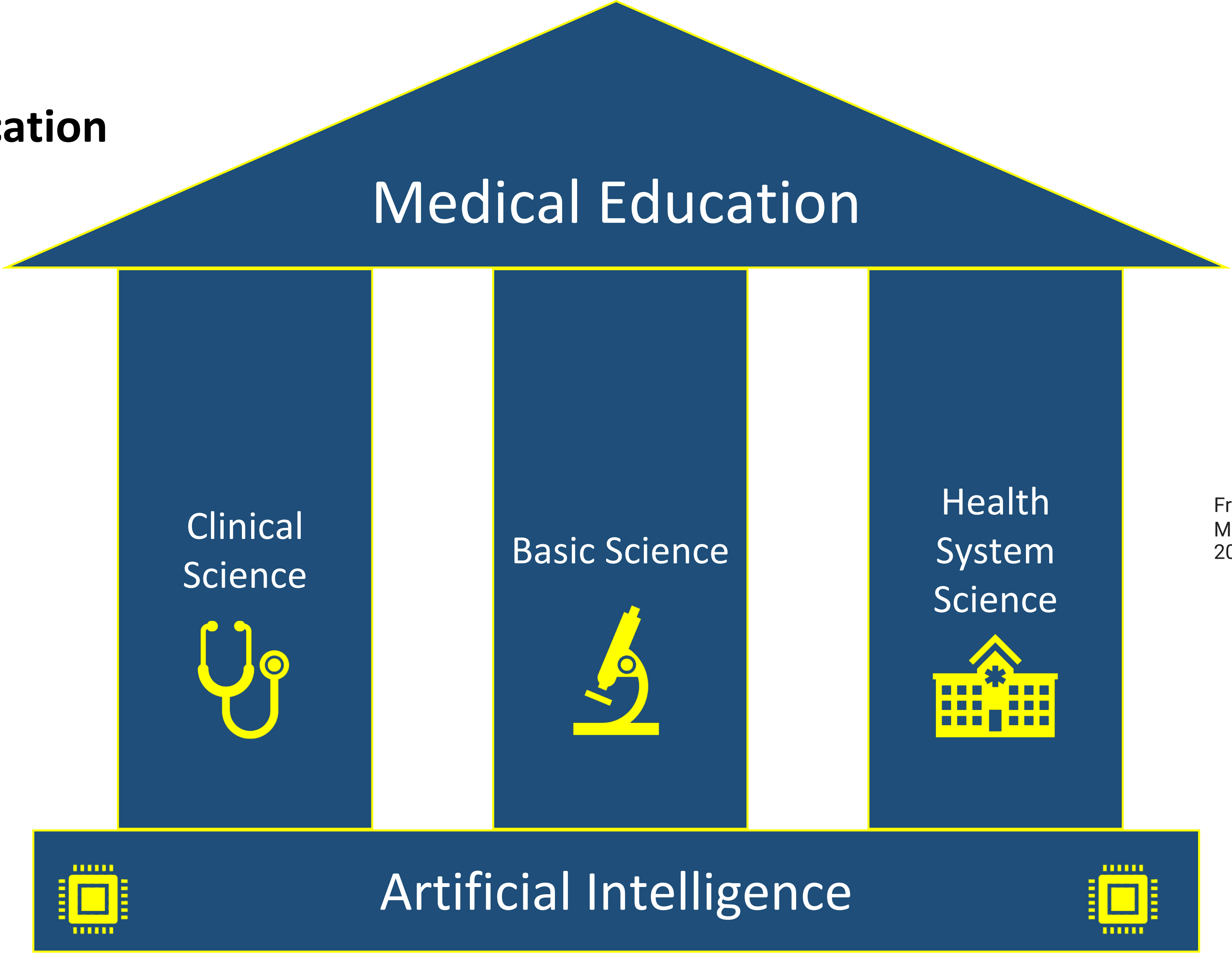
Biotechnomedical (BTM) Model



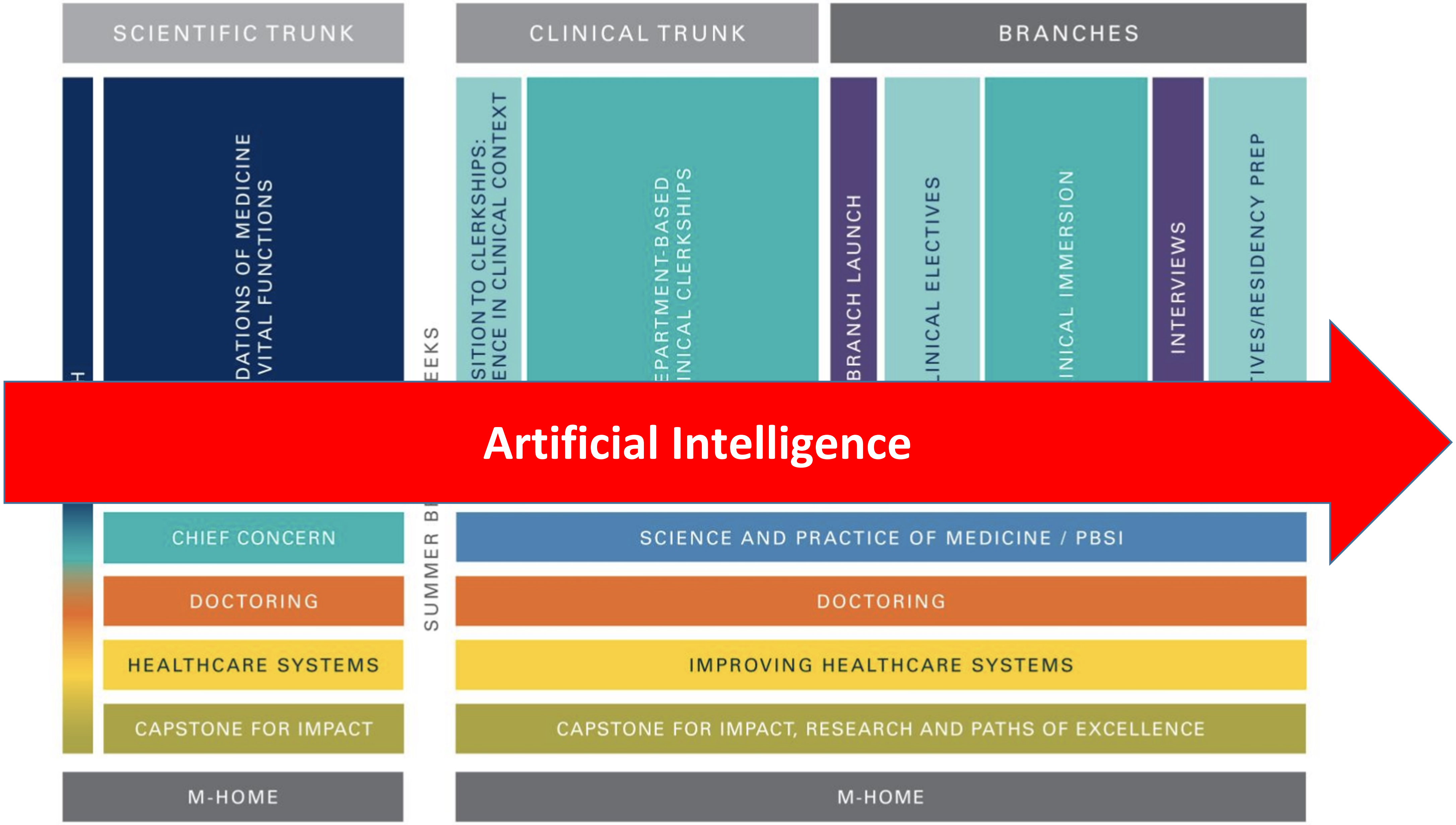
Integration

James CA, Wheelock KM, Woolliscroft JO. Machine Learning: The Next Paradigm Shift in Medical Education. *Acad Med*. 2021;96(7):954-957.

Pillars of Medical Education



Fred HL, Gonzalo JD. Reframing Medical Education. *Tex Heart Inst J*. 2018;45(3):123-125.



SCIENTIFIC TRUNK
CLINICAL TRUNK
BRANCHES
DATIONS OF MEDICINE
VITAL FUNCTIONS
SUMMER B
EEKS
SITION TO CLERKSHIPS:
ENCE IN CLINICAL CONTEXT
EPARTMENT-BASED
INICAL CLERKSHIPS
BRANCH LAUNCH
INICAL ELECTIVES
INICAL IMMERSION
INTERVIEWS
IVES/RESIDENCY PREP
Artificial Intelligence
CHIEF CONCERN
SCIENCE AND PRACTICE OF MEDICINE / PBSI
DOCTORING
DOCTORING
HEALTHCARE SYSTEMS
IMPROVING HEALTHCARE SYSTEMS
CAPSTONE FOR IMPACT
CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE
M-HOME
M-HOME

Biomedical Model

UMMS Scientific Trunk

Foundations of Medicine I: Molecular/Biological Sciences and Evidence-Based Care

Foundations of Medicine II: Host Defense, Physical and Diagnostics and Therapeutics

Current State

CHIEF CONCERN

IMPROVING HEALTHCARE SYSTEMS

DOCTORING

CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE

M-HOME

Biotechnomedical Model

UMMS Scientific Trunk Block 1

Foundations of Medicine I: Molecular/Biological/ Data Sciences and Evidence-Based Care	Foundations of Medicine II: Host Defense, Physical and Technological Diagnostics and Therapeutics

CHIEF CONCERN

IMPROVING HEALTHCARE SYSTEMS

DOCTORING

CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE

M-HOME

Future State?

Biotechnomedical Model Example

UMMS Scientific Trunk Block 6

Foundations of Medicine III: Infection, Hematology, Immunopathology, and Predictive Models

CHIEF CONCERN

IMPROVING HEALTHCARE SYSTEMS

DOCTORING

CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE

M-HOME

Future State?

- **UMMS Block 6**
  - Hematology
  - Infectious diseases
    - Microbes, diagnoses, anti-microbials
    - Sepsis

- **EBM**
  - Critical evaluation of *Epic Sepsis Model* performance

- **Chief Concerns**
  - Integrating output of *Epic Sepsis Model* into clinical reasoning to generate a differential diagnosis

- **Doctoring**
  - Explaining the role of AI/ML (*Epic Sepsis Model*) in decision making

- **Health Systems Science (Improving Health Systems)**
  - Implementing the *Epic Sepsis Model* into the Health System
  - Workflow, regulation, etc.

- **Interprofessional Education**
  - Medical students, CSE students, law students, etc.
    - How could the model be improved?

Research

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

Foundations of Medicine III:
Infection, Hematology, Immunopathology, and
Predictive Models

CHIEF CONCERN

IMPROVING HEALTHCARE SYSTEMS

DOCTORING

CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE

M-HOME

- **UMMS Block 6**
    - Hematology
    - Infectious diseases
        - Microbes, diagnoses, anti-microbials
        - Sepsis

Foundations of Medicine III:
Infection, Hematology, Immunopathology, and
Predictive Models

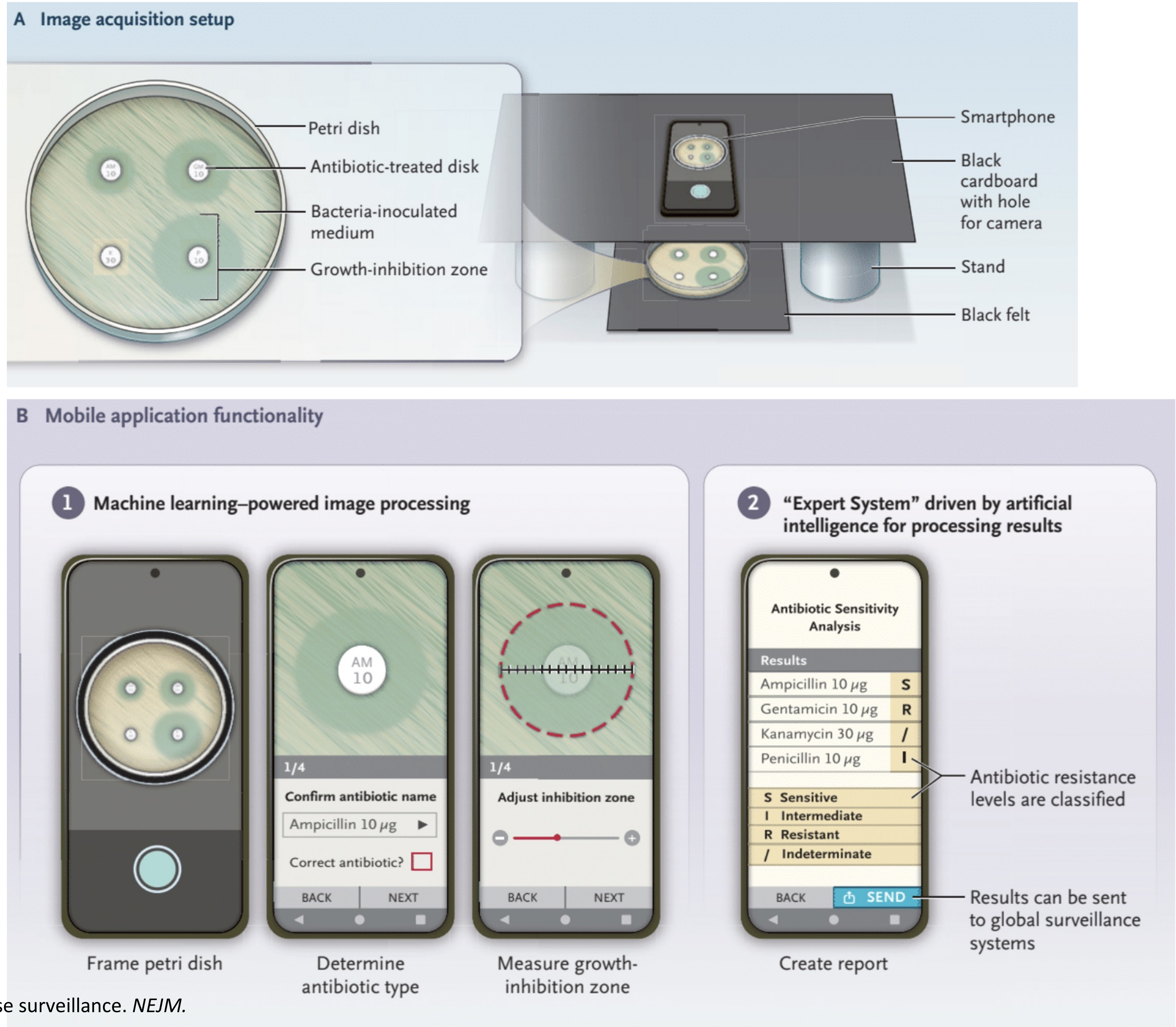
CHIEF CONCERN

IMPROVING HEALTHCARE SYSTEMS

DOCTORING

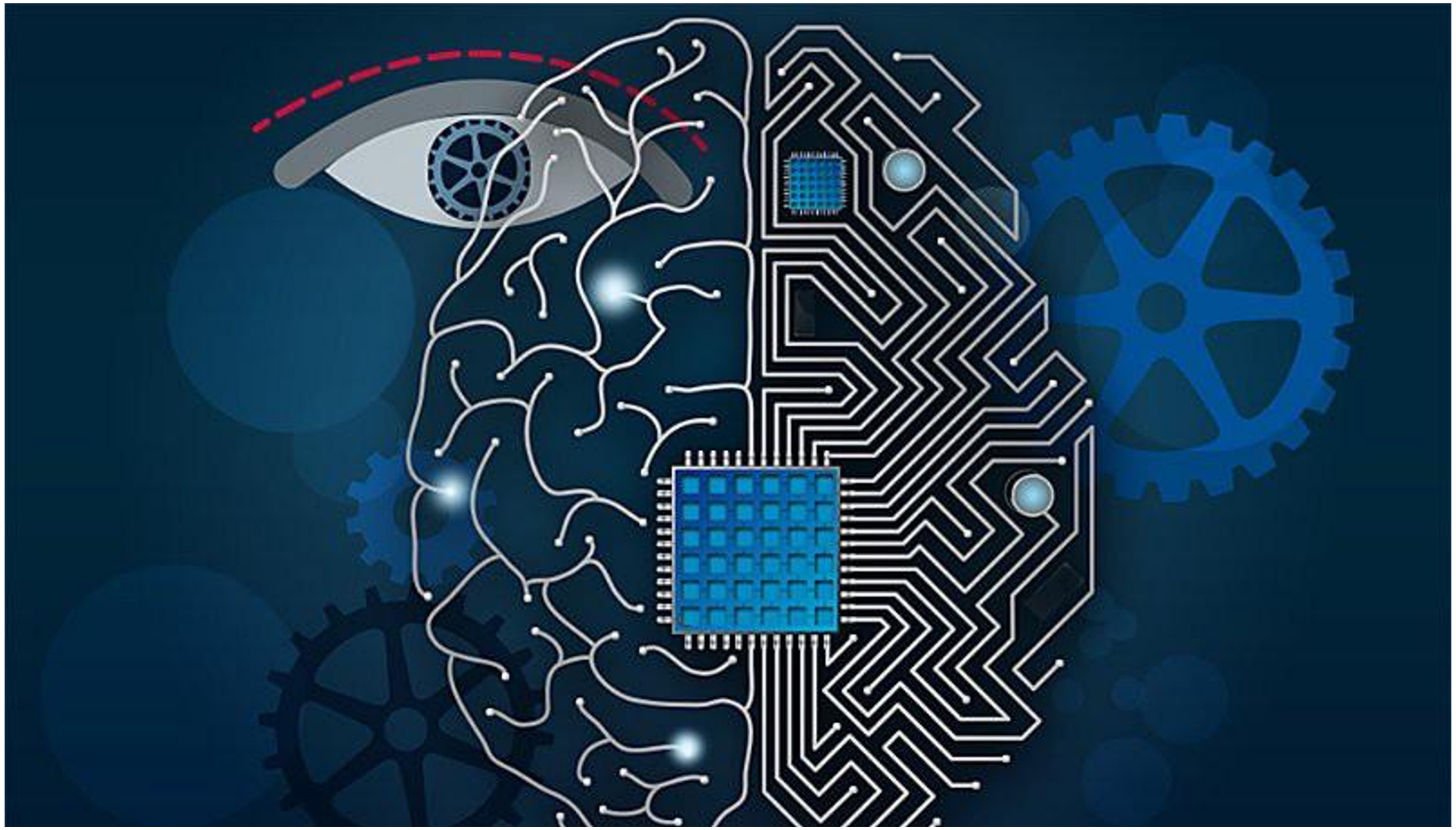
CAPSTONE FOR IMPACT, RESEARCH AND PATHS OF EXCELLENCE

M-HOME



Brownstein JS, Rader B, Astley CM, Tian H. Advances in artificial intelligence for infectious disease surveillance. *NEJM.*

Data Augmented, Technology Assisted Medical Decision Making (DATA-MD)



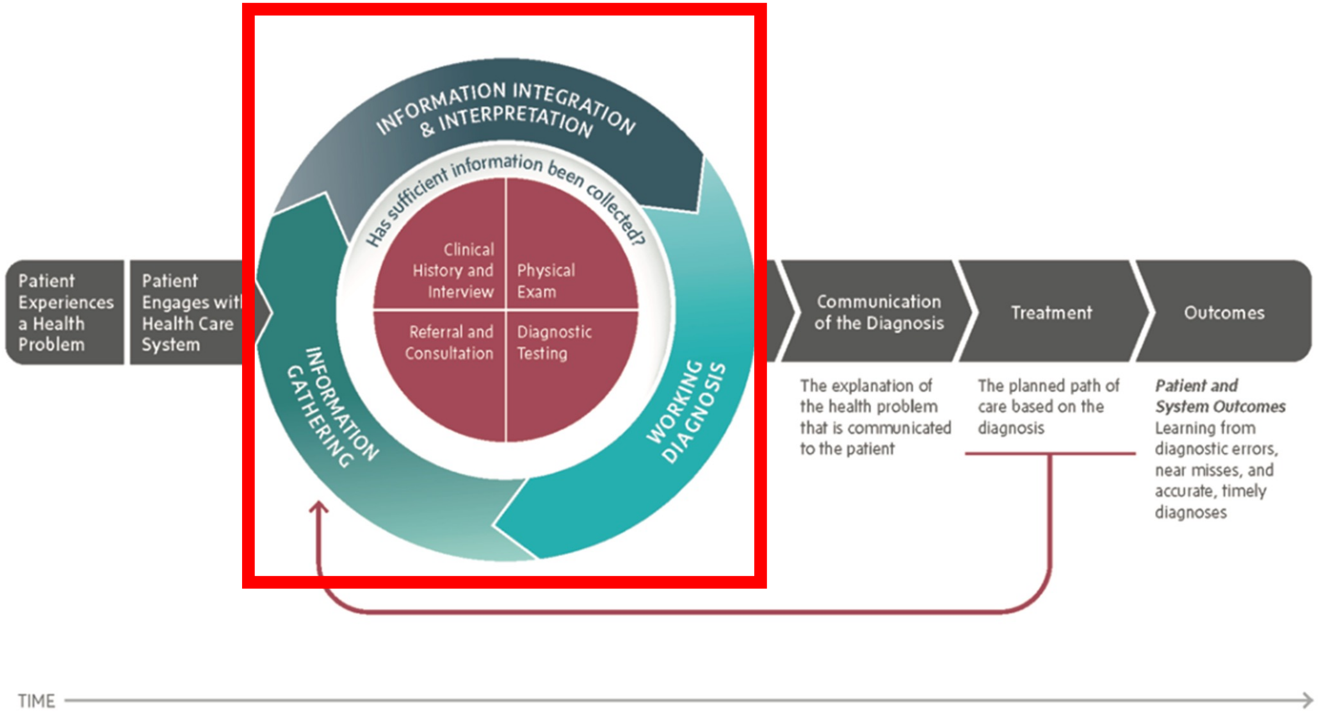
DATA-MD Mission

To develop, implement, and disseminate innovative health care AI/ML curricula that serve as a foundation for medical educators to develop curricula specific to their own institutions and/or specialties.

DATA-MD Team

- Cornelius A. James, MD
- Nancy Allee, MLS, MPH
- Larry Gruppen, PhD
- Benjamin Li (medical student)
- Maggie Makar, PhD
- Brahmajee Nallamothu, MD, MPH
- Nicholson Price, JD, PhD
- Karandeep Singh, MD, MSc
- Jessica Virzi, MSN
- Jenna Wiens, PhD
- James Woolliscroft, MD
- Andrew Wong, MD (U-M House Officer)

DATA-MD and Frameworks



NAM Diagnostic Process

UMMS Evidence-Based Medicine Process

James CA, Wheelock KM, Woolliscroft JO. Machine learning: the next paradigm shift in medical education. *Acad Med.* 2021.96(7): 954-957.

DATA-MD

- Use of AI/ML in diagnostic decision making
  - EBM framework
  - Bayesian approach
- Four online modules
  - Intro to AI/ML in Healthcare
  - Foundational Biostats and Epi in AI/ML for Health Professionals
  - Using AI/ML to Augment Diagnostic Decisions
  - Ethical and Legal use of AI/ML in the Diagnostic Process
- Launch 2023

CENTER FOR ACADEMIC INNOVATION
UNIVERSITY OF MICHIGAN

coursera

GME INNOVATIONS

DATA-MD

- Seven web-based modules
  - Intro to AI in Health Care
  - Methodologies
  - Diagnosis
  - Treatment and Prognosis
  - Law, Ethics, Regulation
  - AI in the Health System
  - Precision Medicine

- Launch 2023

AMA
AMERICAN MEDICAL ASSOCIATION

Additional Curricula

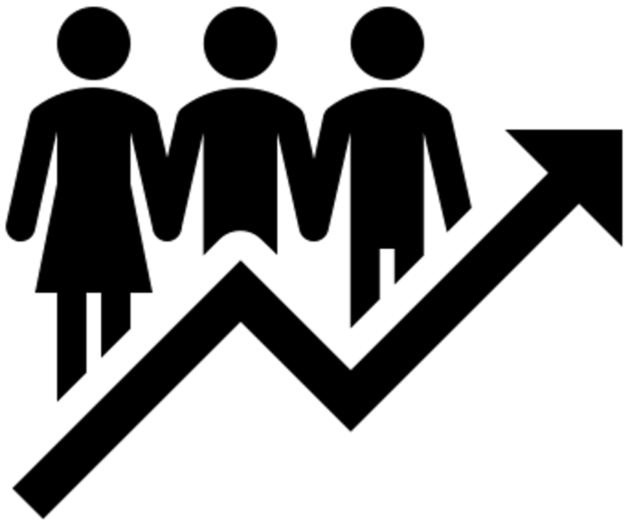
- Five web-based modules
- Foundational
  - Medical students, residents
- Frontline clinicians
  - Brief video series
- 2025

STFM
SOCIETY OF TEACHERS OF FAMILY MEDICINE

GORDON AND BETTY MOORE FOUNDATION

Next Steps

- Curricular review
  - School, course, session level
  - Re-prioritization
- Identify champion(s)
  - Learners, faculty, staff
  - Committees
- Interprofessional collaboration
  - Engage stakeholders
- Faculty development



Take Home Points

- AI/ML in health care is here, and it will continue to march forward with or without physicians.

- AI/ML has the potential to transform the way medicine is practiced.

- Currently, AI/ML instruction in medical education is lacking.
    - We must begin to consider how we incorporate this content into curricula.

- Interprofessional collaboration is essential.